



文献 CSTR:

32001.14.11-6035.csd.2023.0164.zh



文献 DOI:

10.11922/11-6035.csd.2023.0164.zh

文献分类: 信息科学

收稿日期: 2023-11-02

开放同评: 2023-11-06

录用日期: 2023-12-25

发表日期: 2024-03-28

* 论文通信作者

张丽丽: zhll@cnic.cn

数据论文引用计量规律初探——基于中国数据期刊 的案例研究

张丽丽^{1*}, 惠嘉怡¹, 刘瑞霖^{1,2}

1. 中国科学院计算机网络信息中心, 北京 100083

2. 中国科学院大学, 北京 100049

摘要: 数据出版是一种创新数据共享形式。为更好地了解数据出版, 本文通过文献综述和案例调研, 构建了数据论文计量框架, 遴选国内数据期刊《中国科学数据(中英文网络版)》和《全球变化数据学报(中英文)》, 揭示两刊数据出版与数据共享情况。统计涵盖两刊论文指标(学科领域、团队规模、资金来源)、数据指标(数据来源、规格与规模、数据曝光)、引证指标(被引、施引、时间)等30余个特征项。结果显示, 宏观趋势方面, 数据出版成为数据共享的重要途径, 并在规范数据质量、汇聚特色主题数据、推进团队合作、争取资金支持等方面具有优势。中观运营层面, 国家数据中心为数据出版提供平台保障。微观资源层面, 出版数据集兼具学科差异性和跨域共性治理需求。其中, 数据曝光、访问引用转化率等指标, 为数据可复用性测度提供线索。此外, 独立数据出版在我国尚处发展上升期, 建议加强数据文化建设、扩展运营模式、提升资源治理能力与平台建设水平, 完善技术迭代与激励评价机制等。

关键词: 开放数据; 数据论文; 数据出版; 数据重用; 数据引用; 引用计量

引言

开放数据是科研创新的驱动力, 是全球开放科学运动^[1]的核心要素。在我国, 《科学数据管理办法》(2018)^[2]、《“十四五”规划和二〇三五年远景目标》(2020)^[3]等都鼓励数据开放共享。新修订的《中华人民共和国科学技术进步法》(2021)^[4]全文十五处提及“开放”, 强调了“建立健全科学技术资源开放共享机制”(第五十四条), 推动开放科学(第九十五条), 以及公开“科学技术数据”等资源分布和使用情况等内容(第一百零二条)。《中共中央国务院关于构建数据基础制度更好发挥数据要素作用的意见》(2022)^[5]更进一步强调了数据作为新兴生产要素的基本地位, 并将“共享共用”作为推进国家数据基础设施建设的基本原则之一。由此可见, 开放数据的重要性已上升至国家层面。为更好地推进开放数据, 我们有必要追溯共享数据的多维特征, 探索数据共享与重用的紧密关联, 探讨数据潜在价值的生长性, 确保数据开放工作的长期可持续性。

在众多数据共享模式中, 学术出版为开放数据提供了一类途径, 它提升了科研数据的可信度, 增加了学术造假成本, 为更好地重用数据带来可能, 在争夺数据资源、推动科学研究等方面^[6]发挥重要作用。为此, 本文将着重探讨数据论文

的引用计量规律，在传统数据集复用指标，如访问量、下载量等的基础上，借用文献计量学拓展框架，深度揭示学术出版中的共享数据特征与趋势，为数据共享工作提供参考。

1 研究概述

1.1 研究进展

所谓数据出版，即以出版形式推进数据开放重用的过程^[7]，主要包括作为科研附属物的数据出版（如数据附录）、独立数据出版（如数据论文等）以及广义数据出版（即平台数据集发布）三类^[8]。早期学术期刊多以科研附属物的形式规范论文数据。随着数据管理技术革新与数据重用需求的膨胀，数据论文为代表的独立数据出版实践日趋流行，如 Internet Archaeology（1996）、Earth System Science Data（ESSD，2008）、GBIF（2011）、Pensoft Data Journal（2011）、Nature Scientific Data（2014）、全球变化科学研究数据出版系统（2014）、《中国科学数据（中英文网络版）》（2015）等。依附数据平台的广义数据出版，包括机构存储库 PANGAEA（1987; 1995）、arXiv（1991）、RePEc（1992）、PubMedCentral（2000）、DRYAD（2008）、FIGSHARE（2011）、ScienceDB（2015）等实践。与此同时，以机构存储库为代表的广义数据出版也常与前两类数据出版形式关联，即为科研期刊、数据期刊论文出版提供关联数据集的共享服务，保障数据实体共享。这些实践中，以数据论文为代表的独立数据出版，支持数据作者发表数据论文和关联数据集来实现数据共享，值得关注。

所谓数据论文，即数据集的描述性文章^[9-10]，它为面向数据集的引用计量评价^[11-12]提供支持，为更好地评估以数据期刊^[13]为载体的数据集复用提供了线索。其中，数据论文和数据集的引用计量方面，相关研究包括文献计量学指标体系的沿用^[14-18]，数据论文全文计量的引用位置^[19]、引用语义分析，替代计量学对广泛社会影响力指标的吸收^[20-22]，还包括聚焦数据集本身特征^[23-26]等，详见表 1。

表 1 数据集、数据论文的引用计量指标案例

Table1 Examples of citation metrics for datasets and data papers

作者	数据来源	学科领域	数据量	计量框架与指标
陈莉玥等 ^[14]	CSCD 库	全领域	15.45 万余条数据引用记录	引用基础特征、引用数据来源、引用元数据要素及数据类型等定性、定量方法；
焦红等 ^[15]	NCBI PMC 开放获取子集	生物医学	11.24 万篇	被提及数据集数量、被引数据集数量、数据提及频次、数据引用频次、单个数据集被引频次、时间特征、数据集最后更新年份-数据集公开年份，引用常规文献、引用数据论文和引用数据库的频次
张琳等 ^[16]	PLoS Computation Biology	生命科学	2182 条文献记录	学科交叉度 TD
赵蕊菡 ^[17]	ESSD	地球科学	178 篇数据论文	数据论文：被引年代、国别、数据论文对高水平科研成果的影响； 高被引数据论文：自引、引用功能
丁楠等 ^[18]	Web of Science DCI 数据库	人口调查	约 1.38 万条	国别、数据发布量、总被引量、数据平均被引次数、h 指数
卫军朝等 ^[19]	《经济研究》《管理世界》	经济学、管理学、社会学、	911 篇研究论文的数据引用	引用行为：引用位置、引用形式、引用数据类型

作者	数据来源	学科领域	数据量	计量框架与指标
	《社会学研究》《政治学研究》《教育研究》	政治学，以及教育学	行为	数据特征：数据类型（生产形式）、数据来源（国别）、数据作者机构、数据类型（成果形式）、数据来源、数据属性与引用行为相关性等
许鑫等 ^[20]	ESSD	地球科学	489 篇数据论文	潜在影响力：浏览量、下载量和读者数 学术影响力：总被引频次、H 指数和被引对象频次 社会影响力：博客提及量、新闻报道量、推特评论提及量和脸书分享量
翟姗姗等 ^[21]	全球变化科学研究数据出版系统	地球科学	592 篇数据论文/104 篇施引文献	单篇数据论文的学术影响力评价模型：被引对象频次、当年影响因子、下载量三个指标
王雪等 ^[22]	中国知网、Genome Research、BMC Genomics、GENE	生物信息学	1777 篇	科学数据影响力和质量评价模型，包括： 引用指标 ：语种，文献量，数据引用次数，URL，国别，开发者，参考文献类型（总量、URL、功能介绍、第三方文献） 数据影响力 ：数据集个数、文献总被引频词、采用数据集的文档频次、文献总下载频次、均被引频词、均下载频次、文献年均被引频词、年均下载频次
Bai 等 ^[23]	DCI	生命科学与生物医学	13,153 条	数据类型、主题、数据类型/被引率、平均被引次数、高被引数据主题、高被引数据集介绍、数据共享量/数据重用/基金等的相关性显著度检验等
Hood 等 ^[24]	单篇样例	未提及	单篇样例	数据指数：作者已发布的数据集数量（n）（作为主要作者或共同作者）有 n 个或更多数据引用
Park 等 ^[25]	DCI	基因与遗传学	30 个数据集作者、3725 条记录	正式与非正式数据引用
Luo 等 ^[26]	GBIF/CNKI	生命科学	324 篇	施引论文：关键词、年发文量、类型 数据：来源、数据引用形式、应用领域

上述调研可知，数据论文引用计量规律的理论支撑仍可沿袭文献计量学、网络计量学、科学计量学与替代计量学的相关方法。进一步调研指出，论文计量学指标，涵盖期刊平台级、论文级和引用级三个层次。综上，笔者将数据引用计量中有代表性的、潜在测度指标汇总为表 2。

表 2 数据引用计量潜在测度指标^[27-38]

Table 2 Potential indicators for data citation measurement^[27-38]

计量维度		指标分项
期刊		学科领域、期刊名称、期刊类别、期刊影响因子、期刊发文量/频次
论文	出版	题目、出版年份、团队合作（作者数量、机构数量、所属国别/第一作者国别等）、开放获取情况、语种、研究领域、论文类型、论文篇幅、资金来源
	内容	研究设计/研究类型、摘要、论文主题、技术、询证分级、关键词、全文内容分析
数据		数据规模、数据格式、数据来源、数据浏览量、数据下载量
引用	被引	总引用量、篇均引用、篇均他引、年篇均引、引用密度、h 指数、数据指数、引用形式

计量维度		指标分项
	时间	首次被引时长、平均被引时长、首年平均被引（即时指数）、被引半衰期
	施引	施引文献特征（具体指标参见本表格中“论文”计量指标）、高被引论文的单篇分析
	全局	引用关系网络、引用位置、引用语义等

从上述分析可知，数据论文的计量评价，既是衡量数据期刊出版水平的标尺，也为我们更好地探寻依附于数据论文的数据集共享情况提供方法。为此，本研究拟选取国内典型的数据期刊，探究其背后的数据复用规律，重点聚焦：（1）数据论文的特征规律；（2）数据论文关联数据集的特征规律；（3）数据论文的引证复用规律。

1.2 数据来源

本文采取国内较早进行独立数据论文出版的创新型期刊《中国科学数据（中英文网络版）》（以下简称《中国科学数据》，CSDATA）和《全球变化数据学报（中英文）》（以下简称《全球变化数据学报》，GCDATA）作为研究数据来源（详见表3）。数据论文详细信息采集于两期刊官方网站，数据论文发表及引证数据采集于中国知网（www.cnki.net）和Web of Science（www.webofscience.com）。全部数据采集时间为2023年6月1日至2023年8月11日。其中，为完整体现数据出版与重用过程，就知网、Web of Science（WoS）、期刊网站著录数据论文信息等合并取交集，最终用于本文研究的《中国科学数据》数据论文样本量481篇，《全球变化数据学报》数据论文样本量358篇。

表3 遴选数据期刊基本情况

Table 3 Basic information of selected data journals

所属期刊	网站	创刊单位	正式出版时间	发刊频次	发文量（篇） （截至2022年12月31日）	
					总发文量	其中， 数据论文 发文量
《中国科学数据（中英文网络版）》 CN11-6035/N ISSN 2096-2223	http://www.csdata.org/	中国科学院计算机网络信息中心	2016年6月	季刊 线上	510篇	481篇
《全球变化数据学报（中英文）》 CN 10-1493/P ISSN 2096-3645	http://geodoi.ac.cn/	中国科学院地理科学与资源研究所、中国地理学会	2017年3月	季刊 线上/ 印刷	460篇	358篇

1.3 研究方法

数据论文的计量分析，既需要从传统文献计量学指标中寻求规律，又要考量关联数据集的相关特征。基于文献综述相关研究基础，考虑到数据的易获取性、批量数据处理的可行性和研究体量的均衡性，本文遴选文献计量学、网络计量学等有代表性的载文指标，关联数据集的特征描述指标以

及引证指标，构建本文的数据论文计量指标框架，如图 1 所示。考虑到数据论文与数据集的紧密关联，本研究默认数据论文引用反映数据重用情况。

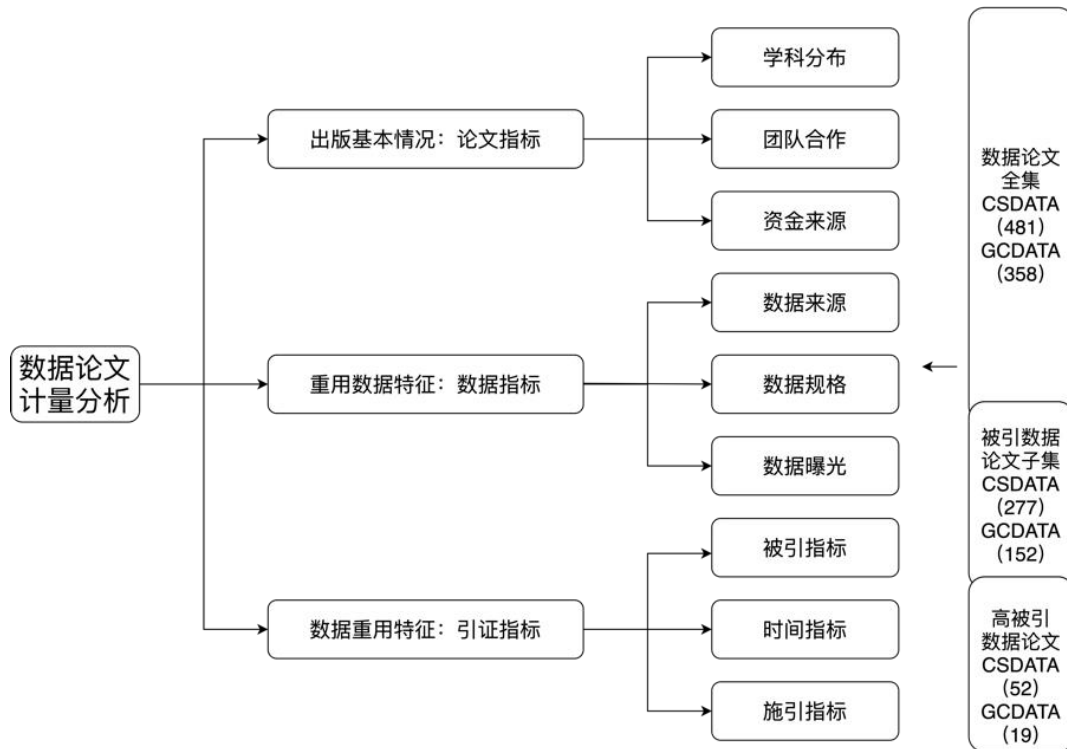


图 1 数据论文计量框架

Figure 1 Citation metrics framework for data papers

为支持上述研究，本论文中的数据采集、处理和分析主要使用 Python 3.7.6 实现，并辅以 R 3.6.3 和网络分析软件 Gephi 0.10.1 进行数据可视化。数据指标计算过程重点使用 pandas、numpy、re、datetime 等 Python 软件包；数据的可视化分析使用 ggplot2 和 ggsankey 等 R 语言包，辅以 matplotlib、seaborn 等 Python 包；数据论文作者合作关系网络分析通过网络可视化软件 Gephi 的 Fruchterman-Reingold 算法绘制。样本数据清洗情况详见表 4。

表 4 数据论文样本

Table 4 Introduction to data paper samples

样本数量（篇）	CSDATA	GCDATA	采集说明
数据论文	481	358	截至 2022 年 12 月 31 日，全文中如无特别说明，即为全样本数据集分析。
被引数据论文	277 (57.59%)	152 (42.46%)	主要用于第 2.3 节数据引证指标分析，数据来源为中国知网（www.cnki.net）和 Web of Science（www.webofscience.com）两刊被引数据之和。
高被引数据论文	52 (10.81%)	19 (5.31%)	主要用于与全样本进行对比使用，根据实际发文被引量分布情况，暂定义“单篇被引用量 ≥ 5 次”为相对高被引论文，简称高被引样本。

2 数据论文共享特征分析

2.1 出版基本情况：论文指标

论文指标重点从数据论文研究背景出发，重点刻画所属学科领域、团队合作情况和资金来源等数据共享工作的重要方面。

2.1.1 学科分布

《中国科学数据》发表全领域数据论文，其中，地球科学（61.33%）、生物学（23.28%）是发刊以来推进数据开放最为活跃的方向（见图2）。此外，开放数据主题还包括信息科学、社会科学、空间天文、物理化学、材料科学及其他。《全球变化数据学报》围绕全球变化主题，以地球科学领域实践为主，考虑到所属学科主题的集中性以及期刊平台本身特征，以下默认全部为地球科学领域数据论文，暂未进一步细分。

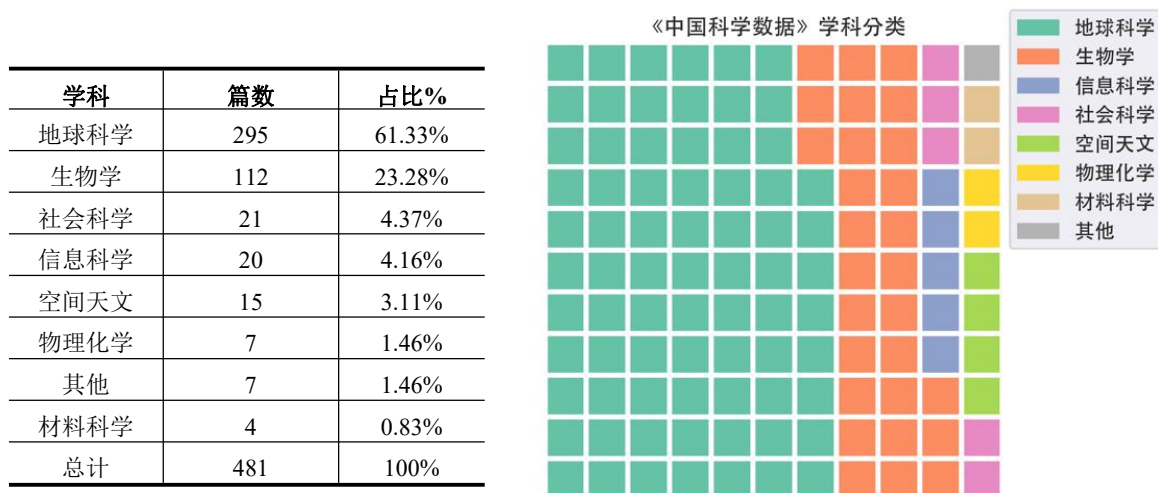


图2 《中国科学数据》数据论文学科分布情况

Figure 2 Disciplinary distribution of data papers in *China Scientific Data*

2.1.2 团队合作

从经验来看，数据集的采集加工与共享重用过程所涉及的技术与领域背景庞杂，需要更多合作支持。统计数据进一步揭示了两刊数据论文的团队合作规模（表5）。《中国科学数据》篇均作者团队规模为5.63人，其中，高被引数据论文篇均作者规模为7.65人；《全球变化数据学报》篇均作者团队规模4.54人，其中，高被引数据论文篇均3.58人。

表5 两刊团队合作规模

Table 5 Collaboration scale in selected journals

期刊	全样本			高被引样本			
	Max.	Min.	Aver.	Proportion	Max.	Min.	Aver.
CSDATA	49	1	5.63	10.81%	44	2	7.65
GCDATA	31	1	4.54	5.31%	9	1	3.58

机构合作网络方面，采用 Gephi 软件绘制数据论文作者所属机构间的合作关系。每个节点代表作者机构，节点颜色区分科研机构所属学科领域，具体由科研机构名称中的关键词加以区分，如包含“生物”“生命”的归为生物学领域，包含“社会”“经济”的归为社会科学领域，以此类推。之后，对自动归类结果辅以人工审核调优。最终形成数据论文合作网络，见图 3。其中，节点大小代表每个机构的平均加权重度，每条边代表一对合作关系，边的粗细代表两机构合作频次高低。

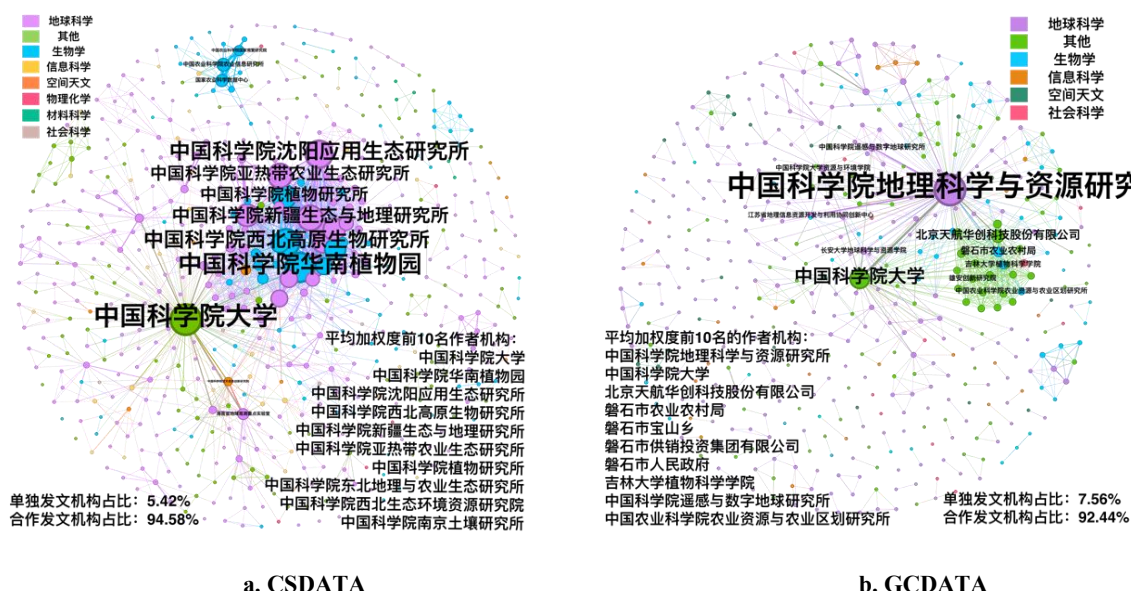


图3 遴选期刊数据论文合作网络

Figure 3 Collaboration network of data papers in selected journals

统计数据显示，数据论文中的跨机构合作日益成为主流。《中国科学数据》和《全球变化数据学报》分别有 15.18% 和 29.33% 的数据论文发表来自单一机构。进一步汇总发现，两刊各有 5.42% 和 7.56%（如图 3 所示）的机构，暂无发表在两刊中的合作数据论文。而两刊合作网络中的活跃机构，包括如中国科学院大学、中国科学院地理科学与资源研究所等。此外，还形成了围绕特定领域主题的高频合作团体。《中国科学数据》围绕热点领域、热点区域进行专题数据论文组稿，例如“中国区域陆地生态系统碳氮水通量及其辅助参数观测专题”、“中国通量观测研究网络（ChinaFLUX）专题”、“中国生态系统研究网络（CERN）专题”等。机构间通力合作推进了数据论文的大批生产，其中围绕科研基础设施“中国生态系统研究网络（CERN）专题”的《2005~2015 年 CERN 光合有效辐射数据集》一文，达到该刊单篇数据论文历史最大合作规模，由 20 家机构 40 个观察台站 44 名研究人员共同署名，关联数据集的署名作者则高达 112 名。

此外，数据出版团队的组建常与特定主题或特定区域关联。例如，围绕特定团队协作，《中国科学数据》生产了“新疆黑蜂蜂蜜”、“新疆云杉”、“宁夏枸杞”等一批特色数据产品。与之类似，《全球变化地球学报》围绕特色主题（如“地理特色产品”、“生态环境保护与可持续发展”等）、特色生境（“黄土高原”、“青藏高原”、“青海湖流域”）、特定技术团队（如国家基础地理信息中心等单位就“全球地表覆盖 30 m 分辨率遥感制图”的系列合作等）推动了相关主题的数据开放。

2.1.3 资金来源

资金投入是支持数据开放的重要保障。开放数据的全生命周期，如数据的生产、加工、存储、分析、数据出版、数据重用等过程都需要耗费大量物力、人力支持。这里假设，标注基金项目数量越多，获得资金支持渠道越丰富，反之亦然。零基金标注，则视为无直接资金资助的数据论文出版。两刊发表数据论文的资金来源情况见表6。《中国科学数据》刊载数据论文篇均获2.49项基金支持，其中高被引篇章的篇均项目支持达2.73项。《全球变化数据学报》篇均项目支持为2.02项，其中高被引篇章的篇均项目支持达到2.26项。多元化的基金支持，为数据生产与数据价值挖掘提供支撑，释放出推进开放数据共享工作的积极信号。进一步观察可知，高被引样本数据论文倾向获得更多来源的资金支持。

表6 单篇数据论文基金项目数

Table 6 Numbers of grants for data papers

期刊	全样本			高被引样本			
	Max.	Min.	Aver.	Proportion	Max.	Min.	Aver.
CSDATA	20	0 (15, 3.12%)	2.49	10.81%	5	1	2.73
GCDATA	8	0 (16, 4.47%)	2.02	5.31%	6	0 (1, 5.26%)	2.26

而无基金支持的数据出版方面，《中国科学数据》占比3.12%；《全球变化数据学报》占比4.47%，且其中包含一篇高被引论文，约占全部高被引数据论文的5.26%。进一步调查发现，《中国科学数据》所发表的无基金标注数据论文绝大部分来自专题组稿，如黄土高原水土保持专题、国家科学数据中心联合专刊、“海上丝绸之路”海洋环境与新能源数据集专题、中国生态系统研究网络（CERN）专题、岩石显微图像专题、黄河流域生态保护与高质量发展专题、减灾研究历史数据集专题等。这些数据论文虽未标注基金，但其数据出版主题背后的研究工作受到国家科研计划、数据基础设施等的长期支持得以延续。由此可见，数据共享工作存在多元化运营模式。而在《全球变化数据学报》，无基金支持的数据论文出版中，部分来自《世界地理数据大百科辞条》，还包括如洛阳民俗博物馆、洛阳古代艺术博物馆、洛阳市山陕会馆文物保管所、中国科学院地理科学与资源研究所及其合作单位等。这些研究机构具有数据捕获的先天性良好条件，为促进数据共享带来契机。这也意味着，数据出版载体本身，作为一种激励模式，在推动数据共享方面，悄然发挥着重要作用。

为进一步检验基金项目是否显著影响数据论文被引情况，对两刊分别设置“是否有基金项目支持”哑变量fund，进行单因子方差分析。其中，有基金支持数据论文，设置fund=1，反之则fund=0，利用单因素方差分析来检验哑变量fund对被引频次影响的显著性。结果显示（表7），两刊单因子方差分析的p值均大于 α ，无法拒绝原假设，故不能认为因子fund显著，即两刊数据论文是否获得基金资助对被引频次无显著影响。

表7 数据论文的基金单因子方差分析

Table 7 One-way ANOVA analysis of grant

期刊	来源	平方和	自由度	F比	p值
CSDATA	因子 fund	33.861	1	1.176	0.279
	误差 e	13797.951	479	-	-

	总和 T	13831.813	480	-	-
GCDATA	因子 fund	0.008	1	0.000	0.985
	误差 e	8247.780	356	-	-
	总和 T	8247.788	357	-	-

显著性水平 $\alpha = 0.05$

《中国科学数据》各领域发表数据论文的项目资助数平均值均大于 1，即表明不同领域数据出版皆获得多源资金支持。其中，生物学领域数据论文所获基金支持最多，达到 2.71 项/篇，信息科学数据论文篇均基金支持最少，约为 1.60 项/篇（见图 4）。

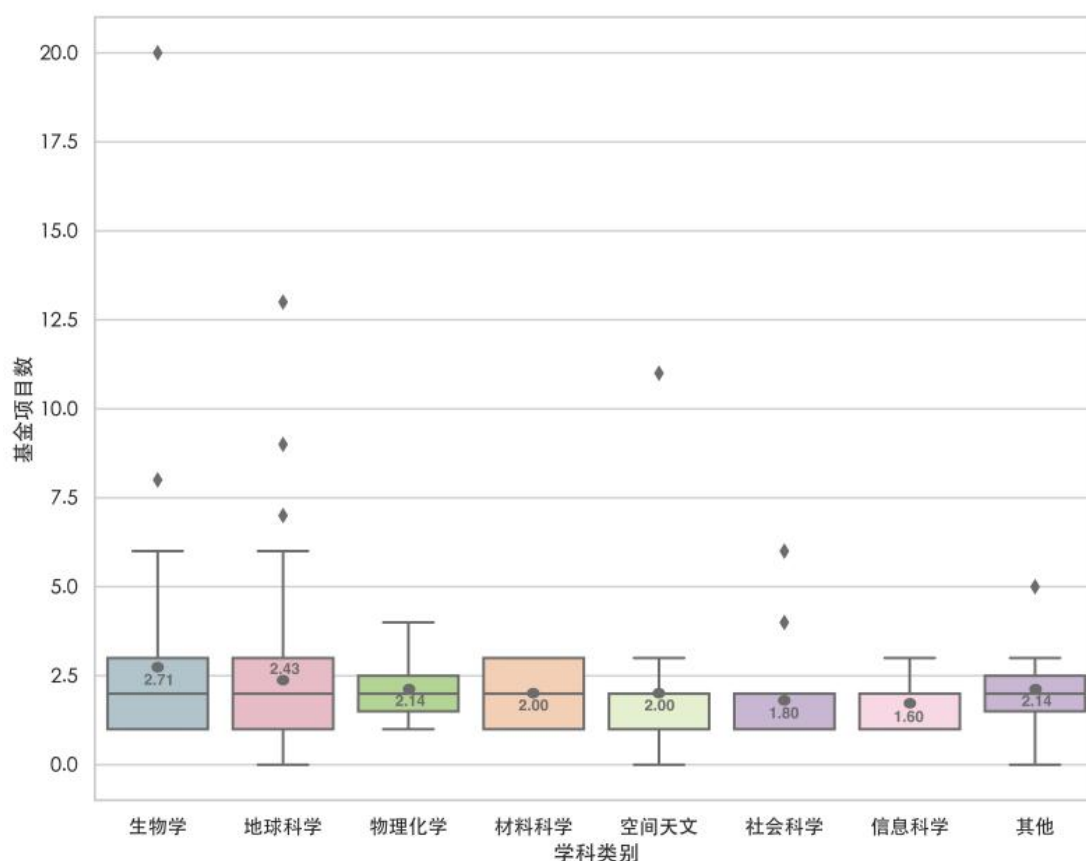


图 4 CSDATA 数据论文分领域基金支持情况

Figure 4 Disciplinary funding support for data papers in CSDATA

2.2 共享数据特征：数据指标

区别于研究论文，数据论文的计量分析同时关注关联数据集的基本特征。根据计量数据的可获取性，以下重点聚焦数据来源、数据规格、数据曝光等计量指标。

2.2.1 数据来源

为保障数据的可获取性，每篇数据论文至少对应一个数据集存储地址。如图 5 所示，根据数据论文 DOI 或数据服务系统网址定位解析可知，《中国科学数据》关联存储库“科学数据银行 ScienceDB”是该刊主要数据存储源。此外，国家级科学数据中心也承担了部分数据存储工作。相关统计显示，

地球科学数据集主要存储在科学数据银行、国家冰川冻土沙漠科学数据中心和国家生态科学数据中心等；生物学数据集多集中于科学数据银行、国家农业科学数据中心和国家生态科学数据中心等；信息科学、社会科学、空间天文等领域数据集则主要存储于科学数据银行 ScienceDB。此外，ScienceDB 也为《全球变化数据学报》提供部分数据存储服务。期刊间的积极互动，在一定程度上展示出数据出版生态建设初露端倪。当然，与《中国科学数据》建立统一标准、允许分散存储的策略不同，《全球变化数据学报》将所出版数据论文极其数据集全部存储于全球变化科学研究数据出版系统。集中化的存储解决方案，为读者提供了一揽子服务，推动数据集和数据论文下载量数据呈现倍增态势。

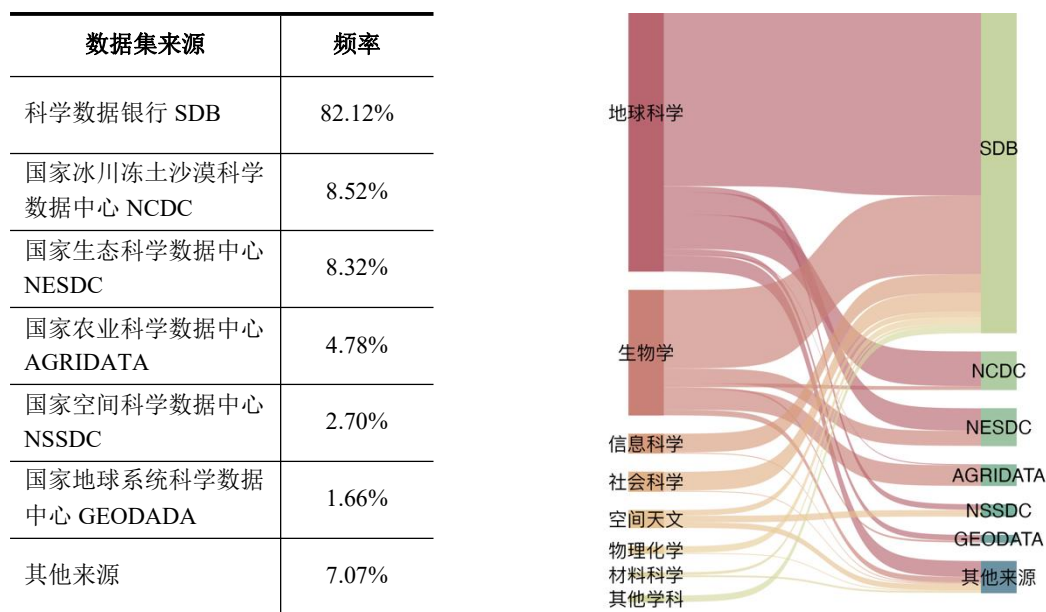


图5 《中国科学数据》数据来源

Figure 5 Sources of data published in *China Scientific Data*

从数据论文作者的角度出发，《中国科学数据》所提供的灵活化方案，为部分具有较成熟数据生产流程与平台的数据生产者出版数据带来便利。考虑到开放数据进程中隐私、安全、知识产权与商业利益保护等场景的存在，受控数据将与开放数据同样流行，标准化的出版与多渠道数据共享的结合，不失为数据期刊的可持续发展良策。

2.2.2 数据规格

数据论文是数据集的使用指南。其中，《中国科学数据》数据论文的平均篇幅约为 10 页，每个数据集平均多于一种数据格式，最流行的数据格式为 xlsx 等。《全球变化数据学报》数据论文平均篇幅约 14 页，每个数据集平均多于一种数据格式，受地球科学学科属性影响，Shapefile 矢量数据居多（详见表 8）。显而易见，学科领域影响着数据格式和数据规模。就《中国科学数据》的进一步调查显示，除 xls、xlsx 等常用办公文档存储格式外，地球科学和生物学领域的图像处理文件格式 tif，地球科学和部分社会科学交叉领域的遥感数据矢量数据 shp 文件，信息科学语料库数据的 txt 纯文本格式，空间天文数据集的 dat 文件，物理化学数据集中的 mdb 数据库文件等，都是与学科或领域软件高度关联的数据存储格式。

表 8 遴选期刊数据论文数据集规格

Table 8 Specifications of data linked to selected data journals

	CSDATA	GCDATA
数据论文长度（单位：页/篇）	10.25	13.75
数据集平均规模（单位：GB）	335.63	0.43
数据集所属格式的平均类型数（类）	1.36	1.75
高频格式（按频率由高到低）	xlsx, tif, shp, xls, jpg	shp, xlsx, tif, kmz, jpg

数据集规模方面（表 9），《中国科学数据》空间天文领域出版包括部分 TB 级规模数据集，如《2011–2017 年“银河画卷”巡天数据集》、《2007–2020 年中国探月工程 VLBI 测量数据集》等；地学遥感、测绘数据集最高逾 50 TB，平均为 375.91 GB；生物学数据集平均规模为 71.70 GB；信息科学数据集平均规模为 12.77 GB，物理化学、社会科学等学科数据集的数量相对较小，常用记录条数来表征。平均数据规模大致描绘出本刊不同学术领域数据论文的关联数据集平均体量，但各学科内所发表数据集的体量仍然千差万别。

表 9 《中国科学数据》数据规模与格式

Table 9 Volumes and formats of datasets in *China Scientific Data*

所属学科	数据规模（GB）			常见数据格式
	Max.	Min.	Aver.	
空间天文	20480.00	1.12×10^{-3}	2887.38	dat, fit, png
地球科学	51200.00	1.62×10^{-6}	375.91	xls, xlsx, tif
生物学	3809.28	1.53×10^{-5}	71.70	xls, xlsx, tif
信息科学	200.00	1.13×10^{-3}	12.77	txt, xls, xlsx, wav
材料科学	20.40	3.42×10^{-3}	5.11	xls, xlsx, json
物理化学	0.10	4.20×10^{-5}	0.02	mdb, xls, xlsx
社会科学	0.07	1.67×10^{-5}	0.01	xls, xlsx, shp
其他	95.00	1.42×10^{-4}	13.58	xls, xlsx, tif

2.2.3 数据曝光

发表数据论文的终极目标即通过引用数据论文重用数据集实体。为此，数据论文及其关联数据集的访问量、下载量值得关注。这里认为，数据论文及其关联数据集的浏览和下载是数据重用的必要非充分条件，即高频数据复用需要较高的浏览量和下载量基数支持，而高浏览量和下载量却不一定带来引用形式的数据重用。考虑到计量数据的可获取性，以及两刊数据实际出版特征，这里默认将两刊官网页面上数据论文浏览量、下载量，视同对应数据集的浏览量、下载量，通过统一口径框定两刊系统统计记录。在此基础上，将浏览量和下载量统称为数据曝光，用以概括数据所受到的广泛关注。其中，定义浏览量表征数据曝光热度，即数据的受关注程度；数据下载量表征数据曝光深度，即数据的重用程度。这是因为，相较于知识获取的便利性（以浏览量为代表的受关注度），数

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/665134113113012112>